Towards automatic detection of reported speech in dialogue using prosodic cues

Towards automatic detection of reported speech in dialogue using prosodic cues

Alessandra Cervone[1], Catherine Lai[1], Silvia Pareti[1,2], Peter Bell[1]

[1]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
[2]Google Inc., 8002 Zürich, Switzerland

s1466724@sms.ed.ac.uk, clai@inf.ed.ac.uk, spareti@google.com, Peter.Bell@ed.ac.uk

Abstract

The phenomenon of reported speech – whereby we quote the words, thoughts and opinions of others, or recount past dialogue – is widespread in conversational speech. Detecting such quotations automatically has numerous applications: for example, in enhancing automatic transcription or spoken language understanding applications. However, the task is challenging, not least because lexical cues of quotations are frequently ambiguous or not present in spoken language. The aim of this paper is to identify potential prosodic cues of reported speech which could be used, along with the lexical ones, to automatically detect quotations and ascribe them to their rightful source, that is reconstructing their attribution relations. In order to do so we analyze SARC, a small corpus of telephone conversations that we have annotated with attribution relations. The results of the statistical analysis performed on the data show how variations in pitch, intensity, and timing features can be exploited as cues of quotations. Furthermore, we build a SVM classifier which integrates lexical and prosodic cues to automatically detect quotations in speech that performs significantly better than chance.

Index Terms: speech recognition, quotations, prosody, attribution relations, discourse, conversational speech.

1. Introduction

Given the pervasiveness of the phenomenon of reported speech in language, automatically identifying the presence of quotations has the potential to improve many automatic speech processing tasks. However, the automatic detection of quotations in speech – compared to written text – is difficult since we cannot rely purely on lexical cues ("he said...", "she told me..." etc), since they are often absent in natural dialogue, and there is of course, no punctuation marking quotation boundaries. As an example of the complexity of this task, consider the following transcribed extract from the Speech Attribution Relations Corpus (SARC) of spoken dialogues analysed in this article:

(1) I said to him when you left do you remember I told you I said to him don't forget Dave if you ever get in trouble give us a call you never know your luck

The presence of the verbs "said" and "told" in the extract suggests that the example may contain reported speech. It is however much harder to determine the number of quotations and their exact boundaries. Some of the possible interpretations of the example could be:

(2) I said to him: "When you left", (do you remember? I told you) I said to him: "Don't forget Dave if you ever get in trouble give us a call". You never know your luck.

(3) I said to him when you left: "Do you remember? I told you" I said to him: "Don't forget Dave if you ever get in trouble give us a call". You never know your luck.

(4) I said to him when you left, (do you remember? I told you) I said to him: "Don't forget Dave if you ever get in trouble give us a call. You never know your luck."

If we rely only on the transcription of Example 1, without listening to the speech itself, even a human would find it very hard to decide which interpretation is the correct one.

Given this complex interaction between lexical and acoustic information, the task of automatically extracting quotations from speech is a very challenging one. However, being able to detect reported speech would be very useful for many applications. Generally, it could improve sentence boundary detection, since quotations could have prosodic marking similar to intonational phrases (which could mislead the system into detecting them as separate sentences). Quotation detection could also improve speaker identification tasks, given that the system could attribute the quoted material to different speakers based on changes in prosodic encoding. Automatically extracting quotations and ascribing them to their source would also be useful for a number of spoken language processing tasks such as information extraction, named entity recognition, and coreference resolution, as well as providing a better understanding of the dialogue structure. Furthermore, being able to rely on the prosodic marking to detect the presence of quotations would be a great advantage for the automatic transcription of speech.

Previous work on reported speech so far has studied this phenomenon mainly from a purely linguistic perspective [1, 2, 3, 4, 5, 6, 7] or considered the prosodic level with only a qualitative approach [8, 9]. While the speech processing literature has investigated prosodic features for automatic punctuation detection, these studies concentrate primarily on full stops, commas and question marks, e.g. [10, 11, 12, 13]. In this work we will try to connect these two approaches to study the potential acoustic correlates of punctuation devices that mark the presence and boundaries of reported speech – such as quotation marks in the case of direct quotations.

This research direction was suggested by the results of our previous study [14], where we performed a linguistic analysis of the quotations in SARC, a small corpus we created to study attribution in speech. SARC was annotated with attribution relations (ARs), i.e. relations holding between the source and the text span containing the reported speech [15]. This annotation followed a scheme adapted from the one developed by [16] for the annotation of PARC, the first large corpus annotated for ARs in written text. The analysis of the ARs in SARC showed that to detect quotations from speech we cannot simply rely only

on lexical information but we need also prosodic information, given that without punctuation the number of potential interpretations of the words becomes exponential (as for example 1). We argued that, in order to build a system able to automatically extract reported speech from spoken language, we need to integrate the lexical cues with prosodic ones. In particular, a pilot acoustic analysis of the ARs in SARC suggested that the quotation span (difficult to detect given its fuzzy boundaries) could be prosodically marked.

In this paper, we test the hypotheses suggested in our previous study by performing an acoustic analysis of the reported speech found in SARC. That is, we test whether prosodic information can be used to automatically detect reported speech. In Section 2, we describe the challenge of detecting ARs in speech and the potential applications of a system able to automatically detect reported speech from spoken language. We also discuss the findings reported in the previous literature regarding the prosodic correlates of reported speech. In Section 3, we provide further details regarding the process of construction and annotation of the corpus and we describe the setup of our experiments. In Section 4, we report the results of the statistical analysis we performed on SARC that show that prosodic correlates of reported speech do exist. We use these features along with lexical ones to train classifiers for detecting reported speech segments and their boundaries – the results suggest that prosodic features can be used successfully for this task.

2. Attribution Relations in Speech

Recovering attribution relations from an automatic speech transcription is hard: even assuming a perfect lexical transcription, identifying quotations in speech using just lexical information is more difficult than in written text due to the presence of disfluencies, such as repetitions, false starts, fillers, sudden changes in the topic or in the structure of the sentence. These problems could deeply affect the understanding of the boundaries and elements of the ARs in the text.

However, the choice of the interpretation leads to very different reconstructions of the structure and meaning of the text. As we saw from the interpretations in Examples 2, 3 and 4, the number of quotations (two in 2 and 3 and one in 4) and the same text span detected as reported speech may vary greatly. Moreover, the same quotation could have different ending boundaries (the sentence "You never know your luck" is considered as reported speech only in 4). Furthermore, the possibility that there could be disfluencies such as repetitions and false starts, leads to the interpretation in 4 (where the two instances of "I said to him" refer to the same attribution event). The correct reconstruction of quotations has consequences for coreference relations (the first "you" in 2 has a different referent to the one in 3) and for our understanding of the time structure (the question "do you remember" is set in the present in 2 and in the past in 3) of the information in the text.

Another complication is given by the fact that sometimes the lexical cue may not be present at all. In this case, without non-lexical information, and particularly prosodic information, we could not hope to achieve the correct interpretation. These cases are not rare. In fact, in [14] we found that lexical cues were missing in about 10% of the quotations in SARC. So, without a better understanding of the relation between the prosodic and lexical encodings of reported speech, the task of automatic attribution extraction from speech seems infeasible.

Past work on the prosodic aspects of quotations in speech has suggested that variations in acoustic features could be considered as correlates of quotations. In particular, these studies found that shifts in pitch [9, 17, 18, 19], intensity [9, 19], and timing features – in particular, pause durations [19] – could act as markers of reported speech. However the literature does not offer a comprehensive insight on the phenomenon of attribution. To our knowledge, the first paper suggesting the possibility of studying the prosodic encoding of reported speech is [9], which presented only a qualitative analysis. Other studies which analyzed prosodic correlates of quotations [17, 18, 19] report statistically significant results, but focus only on one particular aspect of the phenomenon, for example considering only direct reported speech [18, 19] or variations in pitch [17, 18], with some of them relying on rather small corpora [17]. Moreover, these studies provide only descriptive analysis, and have not extended their work to building classifiers to test their findings in a practical setting. To the best of our knowledge, there has been no attempt in the literature to integrate lexical and acoustic features to automatically detect quotations from speech.

In this study we aim to offer a more comprehensive insight on the prosodic dimension of attribution. Using a larger corpus, we examine both direct and indirect reported speech, considering an expanded set of prosodic and timing features based on those investigated in past studies. We also explore the possibility of integrating lexical cues (extracted in our previous findings [14]) with acoustic features to automatically classify reported speech.

3. Experimental Setup

3.1. The Speech Attribution Relations Corpus (SARC)

The Speech Attribution Relations Corpus (SARC) analysed in this study includes 4 informal telephone conversations between 8 english speakers (7 females 1 male). The conversations had a mean duration of 15 minutes (each of the two participants was recorded on a different track). The total duration of the recording is thus about 2 hours (1019 speaker turns). The annotation scheme is described in detail in [14] and [20].

The corpus was manually transcribed at the word level and precise word timings were obtained through Viterbi forced alignment using an automatic speech recognition system trained on a larger corpus of similar material. The dialogues were then annotated for ARs in TranscriberAG [21, 22] by an annotator who had been trained in ARs annotation on the PARC corpus [16]. In total, 209 reported speech segments were identified (125 of the direct type, 84 of the indirect type).

3.2. Classification Tasks

For this study, our goals were to determine how reported speech segments differ prosodically from non-reported speech, and whether these features can be used in classification tasks. We look at features calculated over utterances and words. For the former, we investigate whether we can distinguish reported speech segments from turns containing no reported speech (RS), turns containing some reported speech vs no reported speech (Turn), Direct vs Indirect reported speech (DvI), and Indirect vs Non-Reported Speech (IvNRS). At the word level, we attempt to identify whether a word is the first or last word of a reported speech segment (RS-start, RS-end), looking at turns known to contain at least some reported speech.

Dep. Var.	Sign	Features
RS	+	F0 range, F0 slope, internal pause
	-	Intensity range, Intensity slope
Turn	+	Duration
DvI	+	F0 range
IvNRS	-	Intensity range, intensity slope

Table 1: *Significant effects from utterance level logistic regression models. We compare reported and non-reported speech segments (RS), detection of turns containing some amount of reported speech (Turn), direct vs indirect RS segments (DvI), indirect vs non-RS segments (IvNRS).*

3.3. Feature Extraction

Given the findings of the previous literature and the preliminary analysis performed in [14] we decided to focus on timing (pauses) and prosodic features (pitch and intensity). For the boundary classification tasks we additionally compare the prosodic features with simple lexical cues based on analysis of the SARC and PARC corpora.

Prosodic Features. F0 and intensity data was extracted using Praat at 10ms intervals with linear interpolation and octave jump removal [23]. For F0, parameter settings were automatically determined using the method described in [24], calculated over turns. The F0 values were normalized into semitones relative to speaker mean F0 value (Hz) for that conversation. Intensity measurements were normalized by subtracting the speaker mean for the conversation. We then calculate aggregate statistics — mean, standard deviation, range and slope — over words, turns and reported speech segments. We also consider differences from previous and next word values (Δ_p, Δ_n resp.).

Timing Features. In addition to prosodic features we measure segment durations and gaps to/from the next segment. For utterances, we measure the total amount of internal pausing and the number of words.

Lexical Features. Since reported speech can come with an explicit cue, we look at the efficacy of particular words for predicting these segments. Specifically we include indicators for the top two words preceding ('said', 'thought') and following ('and', 'but') reported speech segments, inclusion in the top 20 preceding and following words in the SARC (p.start20, n.end20), and the top attribution cues from PARC (p.parc, [20]).

4. Results

In the following, we report results of a statistical analysis of prosodic features in reported speech and further classification experiments examining the separability of the various classes outlined above, and comparing this to models including lexical features. For these experiments, features were z-score centred and scaled (with parameters determined over training sets when appropriate).

4.1. Prosodic characteristics of reported speech

We used multilevel logistic regression [25] to get an idea of prosodic variation in our classes. We exclude standard deviation and number of words which are highly correlated with range and duration measurements. We include the speaker as a group level effect to account for individual variation in rates of reported speech production. For utterance level comparisons we focus on intrinsic prosodic properties of the utterance: F0 and intensity aggregates, together with its total duration and summed internal pause duration. Table 1 displays prosodic features whose parameter estimates were significantly different from zero ($p < 0.05$) for the utterance level models. For brevity, we simply indicate the sign of the significant effects. The results highlight several prosodic differences between RS segments to non-RS turns (RS): RS segments have greater F0 (but lower intensity) ranges and slopes, and more internal pausing. We also find that direct quotations have greater F0 range than indirect ones (DvI), while indirect quotes differ from non-quotes in intensity (IvNRS). However, the prosodic differences appear somewhat washed out when we compare turns that include some and no reported speech (Turn). Here, we only see that RS bearing turns are somewhat longer.

Dep. Var.	Sign	Features
RS-start versus:		
non RS-start	+	F0 mean, Δ_p int. mean
	-	Δ_p F0 range
Turn start	+	F0 mean, Δ_p int mean,
		Δ_p int. range, Δ_p F0 range
	-	int. mean, int. range, Δ_p F0 mean
Turn medial	+	F0 mean, Δ_p int mean,
		F0 range, Δ_p int. range
	-	Δ_p F0 range, Δ_n F0 mean
RS end verus:		
non RS-end	+	F0 range, int. range, Δ_n F0 range,
		Δ_p int. range, Δ_n int. mean
	-	Δ_n int. range
Turn end	+	F0 range, Δ_n int. range
		Δ_n F0 mean, Δ_n int. mean,
	-	Int range, Δ_n F0 range
Turn medial	+	F0 range, int. range, Δ_n F0 range,
		Δ_p int range, Δ_p F0 mean
	-	Δ_n int. range, Δ_n F0 mean

Table 2: *Significant boundary effects (logistic regression)..*

The lack of prosodic differences in the turn-to-turn comparison suggests we need to look more closely at the RS segment boundaries. Here, we also consider contextual difference features (Δ_p, Δ_n). We perform three comparisons to RS start words (similarly end words): with all non-RS start words, non-RS turn starting words, and non-RS turn medial words (see Table 2). In general, it appears that RS start words generally have a higher mean F0 and an increased mean intensity from the previous word. This might suggest that RS start boundaries may be similar to intonation phrase resets. However, we also see differences between RS starts and turn initial words with the former having lower intensity mean and range, but higher mean F0. Similarly, the results suggest that RS start words have different prosodic characteristics to turn medial words even though the majority of RS starting points are turn medial (86%).

Differences are also evident with non-RS turn medial and turn ending words compared to RS-end words. Overall, RS-end words display greater F0 range and reduced intensity range compared to the following word. More RS-end points coincide with turn ends (66%), so we expect greater similarity in this case. However, prosodic differences are still evident: RS-end words have smaller intensity range and increased F0 range compared to non-RS turn ends. They also have larger intensity range and mean, and F0 mean compared to the following word.

4.2. Classification Experiments

The statistical analysis above suggests that RS segments have different prosodic characteristics to non-RS segments over their duration and at their boundaries. However, we would also like

Features	RS	Turns	DvI	IvNRS
all	0.72	0.77	0.67	0.71
signif. effects	0.68	0.78	0.62	0.64
FS-all	**0.76**	**0.80**	**0.71**	**0.73**

Table 3: *AUROC results for utterance level classification experiments. FS indicates that forward feature selection was applied.*

Task	Features
RS	Int. slope, int. range, dur, n.words, intern.pause, int SD
Turn	num. words, intern.pause, int. mean, F0 sd
DvI	F0 mean, F0 range, int. range, F0 slope, n.words, dur
IvNRS	Int. slope, int. range, int. SD, dur, n.words
RS start	p.start20, p.gap, int. range, $\Delta_p$int. range, int SD, $\Delta_n$int. mean
RS end	n.gap, Δ_pF0 range, n.end20, $\Delta_n$int. mean, int. range

Table 4: *Best sets from forward feature selection.*

to know whether these differences are enough to automatically detect reported speech. To do this we perform classification experiments using prosodic features. For these experiments, we report 10 fold cross-validation results using SVMs with RBF kernels [26]. For each binary classification task, we report Area Under Reciever Operating Characteristic (AUROC) [27] in order to get an idea of the separability of classes over a range of potential classification thresholds. Using the RBF kernel generally improved on linear kernel and logistic regression results. Because of the unbalanced nature of the data, we found it useful to downsample in the training folds except in the direct vs indirect task (which is more balanced). We also employed forward feature selection (FS) based on AUROC to identify predictive features.

Table 3 presents the results for our utterance level classification tasks. In general, we see that prosodic features perform well above the chance baseline at separating reported and non-reported speech, as well as direct vs indirect quotations. This provides more evidence for the hypothesis that reported speech is prosodically marked. Using only significant effects in the previous statistical analysis provides better performance than the full feature set. However, our best results come from forward feature selection (FS-all). Overall, the selected features (Table 4) include the relevant statistically significant feature sets, as well as correlated features excluded from the statistical analysis.

For the boundary detection task, we also consider the predictiveness of additional lexical and non-lexical contextual features described in Section 3.3 (Table 5). The full non-lexical feature set adds timing features (i.e. previous/next pause) to the prosodic feature set. As for the utterance level classification, the results show that prosodic/non-lexical features can be used for both boundary detection tasks, with large improvements after feature selection, although prosodic separability is much better at RS end points. Lexical features are the strongest cues for the RS starting boundaries, suggesting that lexical cues make prosodic variability within class more acceptable. The best results come from feature selection on the combined set of lexical, timing, and prosodic features (Table 4). The additional lexical and timing features seem to make the significant features from the statistical analysis redundant. In fact, those features provide much less discriminability compared with forward feature selection on prosodic features alone. Overall, we see that prosodic features are useful for the boundary detection task, but their relationship to lexical content could be better exploited, perhaps by using non-linear combinations of these features. In general, we expect that more abstract feature representations that combine lexical and non-lexical information will be beneficial for this task.

Features	RS-start	RS-end
All	*0.82*	*0.86*
Lexical	0.77	0.78
Non-lexical	0.53	0.80
prosody	0.56	0.78
signif. prosody	0.54	0.75
FS-all	**0.86**	**0.92**
FS-Lex	0.77	0.78
FS-non-lexical	0.63	0.89
FS-prosody	0.60	0.80

Table 5: *AUROC for boundary detection tasks.*

5. Discussion and Conclusions

Our study confirmed that quotations have a different prosodic encoding compared to non-reported speech. We also showed how the cues from the acoustic level can be integrated with lexical cues to detect quotations and their boundaries in a spoken corpus. We confirm F0 range differences between reported and non-reported speech. However, we also find intensity features to be important markers of quotations. In fact, we found that indirect quotes are prosodically different to non-reported speech primarily in intensity (contra [17]). Overall, this suggests that prosody adds something more than textual quotation marks.

The results of our experiments show the importance of combining the lexical and prosodic levels in order to extract reported speech from a spoken corpus. They also suggest that quotations in speech have rather different lexical encoding compared to written language. A likely reason that the PARC corpus verbal cues were less effective than those induced from the spoken dialogue data is because of the overlap between some of the top quotative verbs (e.g. 'say', 'thought'). Moreover, other frequent cues in speech, such as 'go' and 'like' [28], are rare in the formal writing style found in the PARC corpus.

We expect that using more lexical information (e.g. the preceding context, verb class information) will further improve results for start boundary detection. Similarly, importing sequence labelling techniques from previous work text attribution should also be helpful [29]. However, in cases where lexical information may be unreliable (e.g. ASR output), or when lexical cues are not present, understanding the prosodic encoding of reported speech is crucial. While we established separability based on pitch, intensity and timing features, further investigation of changes in voice quality (e.g. breathy voice) may also be helpful. Moreover, we expect prosodic markers to help with the detection of more fine grained attribution relations (e.g. reported speech from self vs other). Future work will investigate these issues and look to expanding our current corpus.

6. Acknowledgements

This project was developed as a collaboration between the University of Edinburgh and the University of Pavia, thanks to the Erasmus Placement scholarship. We also wish to thank Bonnie Webber, Irina Prodanof and Tommaso Caselli for their helpful suggestions.

7. References

[1] M. M. Bakhtin, "The dialogic imagination: Four essays by mm bakhtin (m. holquist, ed.; c. emerson & m. holquist, trans.)," 1981.

[2] S. Romaine and D. Lange, "The use of like as a marker of reported speech and thought: A case of grammaticalization in progress," *American speech*, pp. 227–279, 1991.

[3] E. Holt, "Reporting on talk: The use of direct reported speech in conversation," *Research on language and social interaction*, vol. 29, no. 3, pp. 219–245, 1996.

[4] E. Holt and R. Clift, *Reporting talk: Reported speech in interaction*. Cambridge University Press, 2007.

[5] E. Holt, "Reported speech," *The Pragmatics of Interaction: Handbook of Pragmatics Highlights*, vol. 4, pp. 190–205, 2009.

[6] G. Bolden, "The quote and beyond: defining boundaries of reported speech in conversational russian," *Journal of pragmatics*, vol. 36, no. 6, pp. 1071–1118, 2004.

[7] J. Sams, "Quoting the unspoken: An analysis of quotations in spoken discourse," *Journal of Pragmatics*, vol. 42, no. 11, pp. 3147–3160, 2010.

[8] S. Günthner, "Polyphony and the layering of voices in reported dialogues: An analysis of the use of prosodic devices in everyday reported speech," *Journal of pragmatics*, vol. 31, no. 5, pp. 685–708, 1999.

[9] G. Klewitz and E. Couper-Kuhlen, "Quote-unquote? the role of prosody in the contextualization of reported speech sequences," *Universität Konstanz, Philosophische Fakultät, Fachgruppe Sprachwissenschaft*, 1999.

[10] D. Baron, E. Shriberg, and A. Stolcke, "Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues," *Channels*, vol. 20, no. 61, p. 41, 2002.

[11] J.-H. Kim and P. C. Woodland, "A combined punctuation generation and speech recognition system and its performance enhancement using prosody," *Speech Communication*, vol. 41, no. 4, pp. 563–577, Nov. 2003.

[12] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 474–485, Feb. 2012.

[13] J. Kolr and L. Lamel, "Development and Evaluation of Automatic Punctuation for French and English Speech-to-Text." in *INTERSPEECH*, 2012.

[14] A. Cervone, S. Pareti, P. Bell, I. Prodanof, and T. Caselli, "Detecting attribution relations in speech: a corpus study," in *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, 2014, pp. 103–107.

[15] R. Prasad, N. Dinesh, A. Lee, A. Joshi, and B. Webber, "Attribution and its annotation in the penn discourse treebank," *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*, vol. 47, no. 2, pp. 43–64, 2007.

[16] S. Pareti, "A database of attribution relations." in *LREC*, 2012, pp. 3213–3217.

[17] W. Jansen, M. L. Gregory, and J. M. Brenier, "Prosodic correlates of directly reported speech: Evidence from conversational speech," in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[18] R. Bertrand, R. Espesser *et al.*, "Voice diversity in conversation: a case study," in *Proceeings of Speech Prosody 2002*, 2002.

[19] M. Oliveira Jr and D. A. Cunha, "Prosody as marker of direct reported speech boundary," in *Proceedings of Speech Prosody 2004*, 2004.

[20] A. Cervone, "Attribution relations extraction in speech: A lexical-prosodic approach," Master's thesis, University of Pavia, Italy, 2014.

[21] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: a free tool for segmenting, labeling and transcribing speech," in *First International Conference on Language Resources and Evaluation (LREC)*, 1998, pp. 1373–1376.

[22] E. Geoffrois, C. Barras, S. Bird, and Z. Wu, "Transcribing with annotation graphs," in *Second International Conference on Language Resources and Evaluation (LREC)*, 2000, pp. 1517–1521.

[23] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2001.

[24] K. Evanini and C. Lai, "The importance of optimal parameter setting for pitch extraction." *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2291–2291, Oct. 2010.

[25] D. Bates, M. Maechler, B. Bolker, and S. Walker, *lme4: Linear mixed-effects models using Eigen and S4*, 2014, r package version 1.1-7. [Online]. Available: http://CRAN.R-project.org/package=lme4

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[27] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics (Oxford, England)*, vol. 21, no. 20, pp. 3940–3941, Oct. 2005.

[28] I. Buchstaller, "He goes and Im like: The new quotatives revisited," in *The 30th annual meeting on new ways of analyzing variation (NWAV 30)*, 2001, pp. 11–14.

[29] T. O'Keefe, S. Pareti, J. R. Curran, I. Koprinska, and M. Honnibal, "A sequence labelling approach to quote attribution," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 790–799.